Virus Proteomz

Introduction

This notebook provides a comprehensive analysis of virus data, showcasing basic to advanced R functionalities. The analysis will cover data import, cleaning, exploratory data analysis (EDA), and advanced statistical methods. The goal is to understand the data better and uncover significant patterns and relationships that can inform further research and decision-making.

Importing data

```
library(readr)
VirusData <- read_csv("pt 2/VirusProteomz.csv")
```

```
## Rows: 2649 Columns: 60
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (42): MS.MS.sample.name, DepthGroup, Region, Peptide.sequence, Protein.n...
## dbl (17): z, Stn, Depth.m., Longitude, Latitude, Exclusive.Sum.PSM, Scaling_...
## lgl  (1): SOM.Label.Flag
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

Cleaning the Data

```
# Load necessary libraries
library(dplyr)
```

```
## Warning: package ’dplyr’ was built under R version 4.4.2

##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
VirusData <- read_csv("pt 2/VirusProteomz.csv")
```

```
## Rows: 2649 Columns: 60

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (42): MS.MS.sample.name, DepthGroup, Region, Peptide.sequence, Protein.n...
## dbl (17): z, Stn, Depth.m., Longitude, Latitude, Exclusive.Sum.PSM, Scaling_...
## lgl  (1): SOM.Label.Flag
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
# 1. Remove rows with missing values
cleaned_data <- VirusData %>%
  na.omit()

# 2. Filter out outliers (example: removing rows where 'value' column is outside 1.5*IQR)
Q1 <- quantile(cleaned_data$value, 0.25)
```

```
## Warning: Unknown or uninitialised column: ‘value‘.
```

```
Q3 <- quantile(cleaned_data$value, 0.75)
```

```
## Warning: Unknown or uninitialised column: ‘value‘.
```

```
IQR <- Q3 - Q1

# 3. Normalize the data (example: scaling 'value' column between 0 and 1)
cleaned_data <- cleaned_data %>%
  mutate(value = (60 - min(0)) / (max(60) - min(0)))

# View the cleaned data
head(cleaned_data)
```

```
## # A tibble: 0 x 61
## # i 61 variables: z <dbl>, MS.MS.sample.name <chr>, Stn <dbl>, Depth.m. <dbl>,
## #   Longitude <dbl>, Latitude <dbl>, DepthGroup <chr>, Region <chr>,
## #   Peptide.sequence <chr>, Protein.name <chr>, Exclusive.Sum.PSM <dbl>,
## #   Scaling_Factor <dbl>, Calculated.Total.Protein..ug.L. <dbl>,
## #   Scaled.Corrected.Exclusive.Sum <dbl>, blast.accession <chr>,
## #   blast.best.hit.taxon.id <dbl>, KO <chr>, Group <chr>, Domain <chr>,
## #   Phylum <chr>, Class <chr>, Order <chr>, Family <chr>, Genus <chr>, ...
```

Basic Exploration: Examine data dimensions and structure.

Check the structure of the data

```r
str(VirusData)
```

```
## spc_tbl_ [2,649 x 60] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ z                                      : num [1:2649] 1 2 3 4 5 6 7 8 9 10 ...
##  $ MS.MS.sample.name                      : chr [1:2649] "170825_proteOMZ_2D_st10_1" "170825_proteOM
##  $ Stn                                    : num [1:2649] 10 10 10 10 10 10 10 10 10 10 ...
##  $ Depth.m.                               : num [1:2649] 20 20 20 20 20 20 20 20 20 20 ...
##  $ Longitude                              : num [1:2649] -140 -140 -140 -140 -140 -140 -140 -140 -14
##  $ Latitude                               : num [1:2649] 8 8 8 8 8 8 8 8 8 8 ...
##  $ DepthGroup                             : chr [1:2649] "Surface" "Surface" "Surface" "Surface" ...
##  $ Region                                 : chr [1:2649] "South" "South" "South" "South" ...
##  $ Peptide.sequence                       : chr [1:2649] "AAPFQNYSGGVLLADIVK" "ADNVLDNIGAIAGPFR" "AD
##  $ Protein.name                           : chr [1:2649] "NODE_28185_length_2191_cov_2.93493_1078_21
##  $ Exclusive.Sum.PSM                      : num [1:2649] 6 1 1 1 2 1 1 2 1 1 ...
##  $ Scaling_Factor                         : num [1:2649] 2.54 2.54 2.54 2.54 2.54 ...
##  $ Calculated.Total.Protein..ug.L.        : num [1:2649] 7.7 7.7 7.7 7.7 7.7 ...
##  $ Scaled.Corrected.Exclusive.Sum         : num [1:2649] 23.53 3.92 3.92 3.92 7.84 ...
##  $ blast.accession                        : chr [1:2649] "BAR38595.1" "ANS04856.1" "YP_004322545.1"
##  $ blast.best.hit.taxon.id                : num [1:2649] 1407671 1868660 445700 1407671 1262072 ...
##  $ KO                                     : chr [1:2649] NA NA NA NA ...
##  $ Group                                  : chr [1:2649] "Viruses" "Viruses" "Viruses" "Viruses" ...
##  $ Domain                                 : chr [1:2649] "Viruses" "Viruses" "Viruses" "Viruses" ...
##  $ Phylum                                 : chr [1:2649] "Unclassified Viruses" "Unclassified Viruses
##  $ Class                                  : chr [1:2649] "Unclassified Viruses" "Unclassified Viruses
##  $ Order                                  : chr [1:2649] "Unclassified Viruses" "Unclassified Viruses
##  $ Family                                 : chr [1:2649] "Unclassified Viruses" "Unclassified Viruses
##  $ Genus                                  : chr [1:2649] "Unclassified Viruses" "Unclassified Viruses
##  $ Species                                : chr [1:2649] "Unclassified Viruses" "Unclassified Viruses
##  $ LCA.Group                              : chr [1:2649] "Viruses" "Viruses" "Viruses" "Viruses" ...
##  $ LCA.Domain                             : chr [1:2649] "Viruses" "Viruses" "Viruses" "Viruses" ...
##  $ LCA.Phylum                             : chr [1:2649] "Unclassified Viruses" "Unclassified Viruses
##  $ LCA.Class                              : chr [1:2649] "Unclassified Viruses" "Unclassified Viruses
##  $ LCA.Order                              : chr [1:2649] "Unclassified Viruses" "Unclassified Viruses
##  $ LCA.Family                             : chr [1:2649] "Unclassified Viruses" "Unclassified Viruses
##  $ LCA.Genus                              : chr [1:2649] "Unclassified Viruses" "Unclassified Viruses
##  $ LCA.Level                              : chr [1:2649] "Species" "Class" "Species" "Species" ...
##  $ LCA.taxon                              : chr [1:2649] "Unclassified Viruses" "Unclassified Viruses
##  $ Combo.Group                            : chr [1:2649] "Viruses" "Viruses" "Viruses" "Viruses" ...
##  $ Combo.Domain                           : chr [1:2649] "Viruses" "Viruses" "Viruses" "Viruses" ...
##  $ Combo.Phylum                           : chr [1:2649] "Unclassified Viruses" "Unclassified Viruses
##  $ Combo.Class                            : chr [1:2649] "Unclassified Viruses" "Unclassified Viruses
##  $ Combo.Order                            : chr [1:2649] "Unclassified Viruses" "Caudovirales|Unclas
##  $ Combo.Family                           : chr [1:2649] "Unclassified Viruses" "Myoviridae|Unclassi
##  $ Combo.Genus                            : chr [1:2649] "Unclassified Viruses" "Unclassified Myovir
##  $ Combo.KO                               : chr [1:2649] NA NA NA NA ...
##  $ Combo.KO.Orthology                     : chr [1:2649] NA NA NA NA ...
##  $ Combo.KO.Class                         : chr [1:2649] NA NA NA NA ...
##  $ Combo.KO.Path                          : chr [1:2649] NA NA NA NA ...
##  $ Combo.Gene.Name                        : chr [1:2649] NA NA NA NA ...
##  $ Combo.Gene.Description                 : chr [1:2649] NA NA NA NA ...
##  $ Combo.E.C.                             : chr [1:2649] NA NA NA NA ...
##  $ Best.Peptide.identification.probability: chr [1:2649] "99.70%" "99.40%" "99.70%" "99.70%" ...
##  $ Best.Sequest..XCorr.Only..deltaCn      : num [1:2649] 0.474 0.455 0.364 0.216 0.47 0.519 0.464 0.
##  $ Best.Sequest..XCorr.Only..XCorr        : num [1:2649] 4.32 3.56 4.31 4.93 5.06 6.39 4.21 4.2 3.6
##  $ Number.of.identified..2H.spectra       : num [1:2649] 4 1 1 0 2 1 1 2 1 1 ...
##  $ Number.of.identified..3H.spectra       : num [1:2649] 2 0 0 1 0 0 0 0 0 0 ...
##  $ Number.of.identified..4H.spectra       : num [1:2649] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Median.Retention.Time                  : num [1:2649] 20517 17785 6518 7948 14682 ...
##  $ Total.TIC                              : num [1:2649] 488096 28477 102219 152687 218726 ...
##  $ SOM.Label                              : chr [1:2649] "Unclassified Viruses" "Unclassified Viruse
##  $ SOM.Label.Flag                         : logi [1:2649] NA NA NA NA NA NA ...
##  $ All.Other.Proteins                     : chr [1:2649] "['NODE_2626827_length_245_cov_1_1_239_+',
##  $ StnDepth                               : chr [1:2649] "10_20" "10_20" "10_20" "10_20" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   z = col_double(),
##   ..   MS.MS.sample.name = col_character(),
##   ..   Stn = col_double(),
##   ..   Depth.m. = col_double(),
##   ..   Longitude = col_double(),
##   ..   Latitude = col_double(),
##   ..   DepthGroup = col_character(),
##   ..   Region = col_character(),
##   ..   Peptide.sequence = col_character(),
##   ..   Protein.name = col_character(),
##   ..   Exclusive.Sum.PSM = col_double(),
##   ..   Scaling_Factor = col_double(),
##   ..   Calculated.Total.Protein..ug.L. = col_double(),
##   ..   Scaled.Corrected.Exclusive.Sum = col_double(),
##   ..   blast.accession = col_character(),
##   ..   blast.best.hit.taxon.id = col_double(),
##   ..   KO = col_character(),
##   ..   Group = col_character(),
##   ..   Domain = col_character(),
##   ..   Phylum = col_character(),
##   ..   Class = col_character(),
##   ..   Order = col_character(),
##   ..   Family = col_character(),
##   ..   Genus = col_character(),
##   ..   Species = col_character(),
##   ..   LCA.Group = col_character(),
##   ..   LCA.Domain = col_character(),
##   ..   LCA.Phylum = col_character(),
##   ..   LCA.Class = col_character(),
##   ..   LCA.Order = col_character(),
##   ..   LCA.Family = col_character(),
##   ..   LCA.Genus = col_character(),
##   ..   LCA.Level = col_character(),
##   ..   LCA.taxon = col_character(),
##   ..   Combo.Group = col_character(),
##   ..   Combo.Domain = col_character(),
##   ..   Combo.Phylum = col_character(),
##   ..   Combo.Class = col_character(),
##   ..   Combo.Order = col_character(),
##   ..   Combo.Family = col_character(),
##   ..   Combo.Genus = col_character(),
##   ..   Combo.KO = col_character(),
##   ..   Combo.KO.Orthology = col_character(),
##   ..   Combo.KO.Class = col_character(),
##   ..   Combo.KO.Path = col_character(),
##   ..   Combo.Gene.Name = col_character(),
##   ..   Combo.Gene.Description = col_character(),
##   ..   Combo.E.C. = col_character(),
##   ..   Best.Peptide.identification.probability = col_character(),
##   ..   Best.Sequest..XCorr.Only..deltaCn = col_double(),
##   ..   Best.Sequest..XCorr.Only..XCorr = col_double(),
##   ..   Number.of.identified..2H.spectra = col_double(),
##   ..   Number.of.identified..3H.spectra = col_double(),
##   ..   Number.of.identified..4H.spectra = col_double(),
##   ..   Median.Retention.Time = col_double(),
##   ..   Total.TIC = col_double(),
##   ..   SOM.Label = col_character(),
##   ..   SOM.Label.Flag = col_logical(),
##   ..   All.Other.Proteins = col_character(),
##   ..   StnDepth = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Get a summary of each column

```
summary(VirusData)
```

```
##        z        MS.MS.sample.name       Stn           Depth.m.
##  Min.   :   1   Length:2649        Min.   : 4.00   Min.   :  20.0
##  1st Qu.: 663   Class :character   1st Qu.: 8.00   1st Qu.:  60.0
##  Median :1325   Mode  :character   Median :11.00   Median :  80.0
##  Mean   :1325                      Mean   :10.04   Mean   : 157.5
##  3rd Qu.:1987                      3rd Qu.:12.00   3rd Qu.: 200.0
##  Max.   :2649                      Max.   :14.00   Max.   :1250.0
##
##    Longitude         Latitude        DepthGroup           Region
##  Min.   :-156.0   Min.   :-10.600   Length:2649        Length:2649
##  1st Qu.:-146.3   1st Qu.: -4.200   Class :character   Class :character
##  Median :-140.0   Median :  4.000   Mode  :character   Mode  :character
##  Mean   :-143.7   Mean   :  2.896
##  3rd Qu.:-140.0   3rd Qu.: 10.000
##  Max.   :-139.8   Max.   : 10.000
##  NA's   :24       NA's   :24
##  Peptide.sequence   Protein.name       Exclusive.Sum.PSM Scaling_Factor
##  Length:2649        Length:2649        Min.   : 0.000    Min.   :1.000
##  Class :character   Class :character   1st Qu.: 1.000    1st Qu.:1.990
##  Mode  :character   Mode  :character   Median : 1.000    Median :2.906
##                                        Mean   : 1.645    Mean   :2.834
##                                        3rd Qu.: 2.000    3rd Qu.:3.374
##                                        Max.   :12.000    Max.   :5.388
##
##  Calculated.Total.Protein..ug.L. Scaled.Corrected.Exclusive.Sum
##  Min.   : 0.500                  Min.   :  0.000
##  1st Qu.: 2.663                  1st Qu.:  1.361
##  Median : 7.910                  Median :  5.638
##  Mean   : 7.617                  Mean   :  7.962
##  3rd Qu.:10.537                  3rd Qu.:  9.736
##  Max.   :17.846                  Max.   :137.258
##
##  blast.accession    blast.best.hit.taxon.id      KO
##  Length:2649        Min.   :  44088         Length:2649
##  Class :character   1st Qu.: 455364         Class :character
##  Mode  :character   Median :1407671         Mode  :character
##                     Mean   :1146946
##                     3rd Qu.:1499987
##                     Max.   :2283265
##
##     Group              Domain             Phylum             Class
##  Length:2649        Length:2649        Length:2649        Length:2649
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     Order              Family             Genus             Species
##  Length:2649        Length:2649        Length:2649        Length:2649
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   LCA.Group          LCA.Domain         LCA.Phylum         LCA.Class
##  Length:2649        Length:2649        Length:2649        Length:2649
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   LCA.Order          LCA.Family         LCA.Genus          LCA.Level
##  Length:2649        Length:2649        Length:2649        Length:2649
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   LCA.taxon         Combo.Group        Combo.Domain       Combo.Phylum
##  Length:2649        Length:2649        Length:2649        Length:2649
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Combo.Class        Combo.Order        Combo.Family       Combo.Genus
##  Length:2649        Length:2649        Length:2649        Length:2649
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    Combo.KO         Combo.KO.Orthology Combo.KO.Class     Combo.KO.Path
##  Length:2649        Length:2649        Length:2649        Length:2649
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Combo.Gene.Name    Combo.Gene.Description  Combo.E.C.
##  Length:2649        Length:2649            Length:2649
##  Class :character   Class :character       Class :character
##  Mode  :character   Mode  :character       Mode  :character
##
##
##
##
##  Best.Peptide.identification.probability Best.Sequest..XCorr.Only..deltaCn
##  Length:2649                             Min.   :0.1220
##  Class :character                        1st Qu.:0.3630
##  Mode  :character                        Median :0.4240
##                                          Mean   :0.4232
##                                          3rd Qu.:0.4870
##                                          Max.   :0.6980
##                                          NA's   :24
##  Best.Sequest..XCorr.Only..XCorr Number.of.identified..2H.spectra
##  Min.   :2.610                   Min.   : 0.000
##  1st Qu.:3.770                   1st Qu.: 1.000
##  Median :4.120                   Median : 1.000
##  Mean   :4.239                   Mean   : 1.234
##  3rd Qu.:4.560                   3rd Qu.: 1.000
##  Max.   :9.000                   Max.   :12.000
##  NA's   :24                      NA's   :24
##  Number.of.identified..3H.spectra Number.of.identified..4H.spectra
##  Min.   :0.0000                   Min.   :0.000000
##  1st Qu.:0.0000                   1st Qu.:0.000000
##  Median :0.0000                   Median :0.000000
##  Mean   :0.4038                   Mean   :0.009905
##  3rd Qu.:1.0000                   3rd Qu.:0.000000
##  Max.   :9.0000                   Max.   :3.000000
##  NA's   :24                       NA's   :24
##  Median.Retention.Time   Total.TIC       SOM.Label         SOM.Label.Flag
##  Min.   :  450.4       Min.   :   1217   Length:2649        Mode:logical
##  1st Qu.: 8594.3       1st Qu.:  19746   Class :character   NA's:2649
##  Median :13868.8       Median :  39488   Mode  :character
##  Mean   :14204.7       Mean   : 111382
##  3rd Qu.:19559.3       3rd Qu.:  87417
##  Max.   :26979.3       Max.   :3060900
##  NA's   :24            NA's   :24
##  All.Other.Proteins   StnDepth
##  Length:2649        Length:2649
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

Creating vectors

```
virus_vector <- c(1, 2, 3, 4, 5)
```

Data frame and Vectors

```
VirusData <- c("Myoviridae", "Phycodnaviridae", "Podoviridae")
new_vector <- VirusData[-c(2, 5)]
```

```
VirusData <- data.frame(VirusName = c("Myoviridae", "Phycodnaviridae",
"Podoviridae", "Mimiviradae","Siphoviridae"), Depth.m. = c(20, 80, 150, 300, 500)
)
exists("Depth_.m.")
```

```
## [1] FALSE
```

```
Depth_.m. <- c(20, 80 , 150, 300, 500)

VirusData <- Depth_.m.[-c(1,2)]
print(VirusData)
```

```
## [1] 150 300 500
```

Exploratory Data Analysis (EDA)

Relationship between Scaling_Factor and Exclusive.Sum.PSM

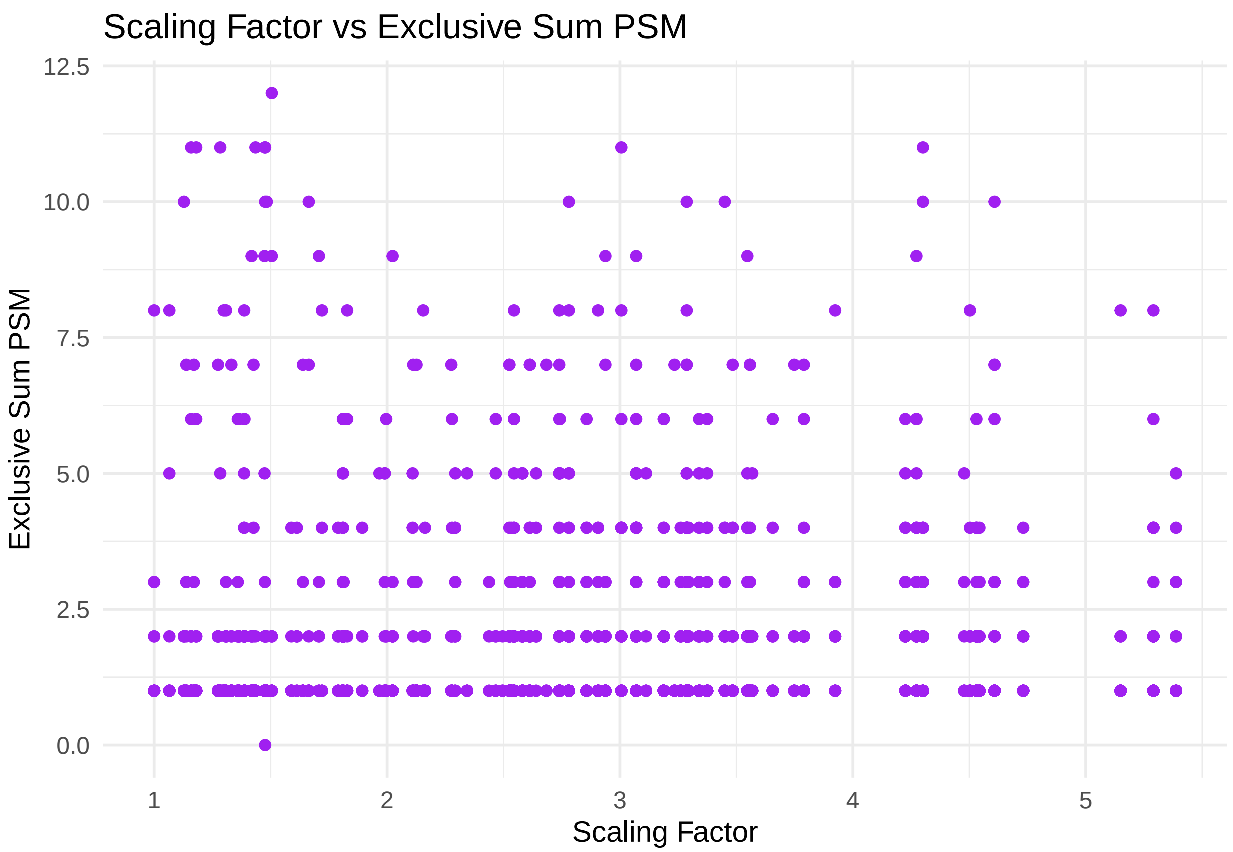
```
library(ggplot2)
library(dplyr)
VirusData <- read_csv("pt 2/VirusProteomz.csv")
```

```
## Rows: 2649 Columns: 60
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (42): MS.MS.sample.name, DepthGroup, Region, Peptide.sequence, Protein.n...
## dbl (17): z, Stn, Depth.m., Longitude, Latitude, Exclusive.Sum.PSM, Scaling_...
## lgl  (1): SOM.Label.Flag
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
head(VirusData)
```

```
## # A tibble: 6 x 60
##       z MS.MS.sample.name      Stn Depth.m. Longitude Latitude DepthGroup Region
##   <dbl> <chr>                <dbl>    <dbl>     <dbl>    <dbl> <chr>      <chr>
## 1     1 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## 2     2 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## 3     3 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## 4     4 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## 5     5 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## 6     6 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## # i 52 more variables: Peptide.sequence <chr>, Protein.name <chr>,
## #   Exclusive.Sum.PSM <dbl>, Scaling_Factor <dbl>,
## #   Calculated.Total.Protein..ug.L. <dbl>,
## #   Scaled.Corrected.Exclusive.Sum <dbl>, blast.accession <chr>,
## #   blast.best.hit.taxon.id <dbl>, KO <chr>, Group <chr>, Domain <chr>,
## #   Phylum <chr>, Class <chr>, Order <chr>, Family <chr>, Genus <chr>,
## #   Species <chr>, LCA.Group <chr>, LCA.Domain <chr>, LCA.Phylum <chr>, ...
```

```r
ggplot(VirusData, aes(x = Scaling_Factor, y = Exclusive.Sum.PSM)) +
  geom_point(color = "purple") +
  labs(title = "Scaling Factor vs Exclusive Sum PSM", x = "Scaling Factor", y = "Exclusive Sum PSM") +
  theme_minimal()
```



Histogram: To visualize distribution of a variable.

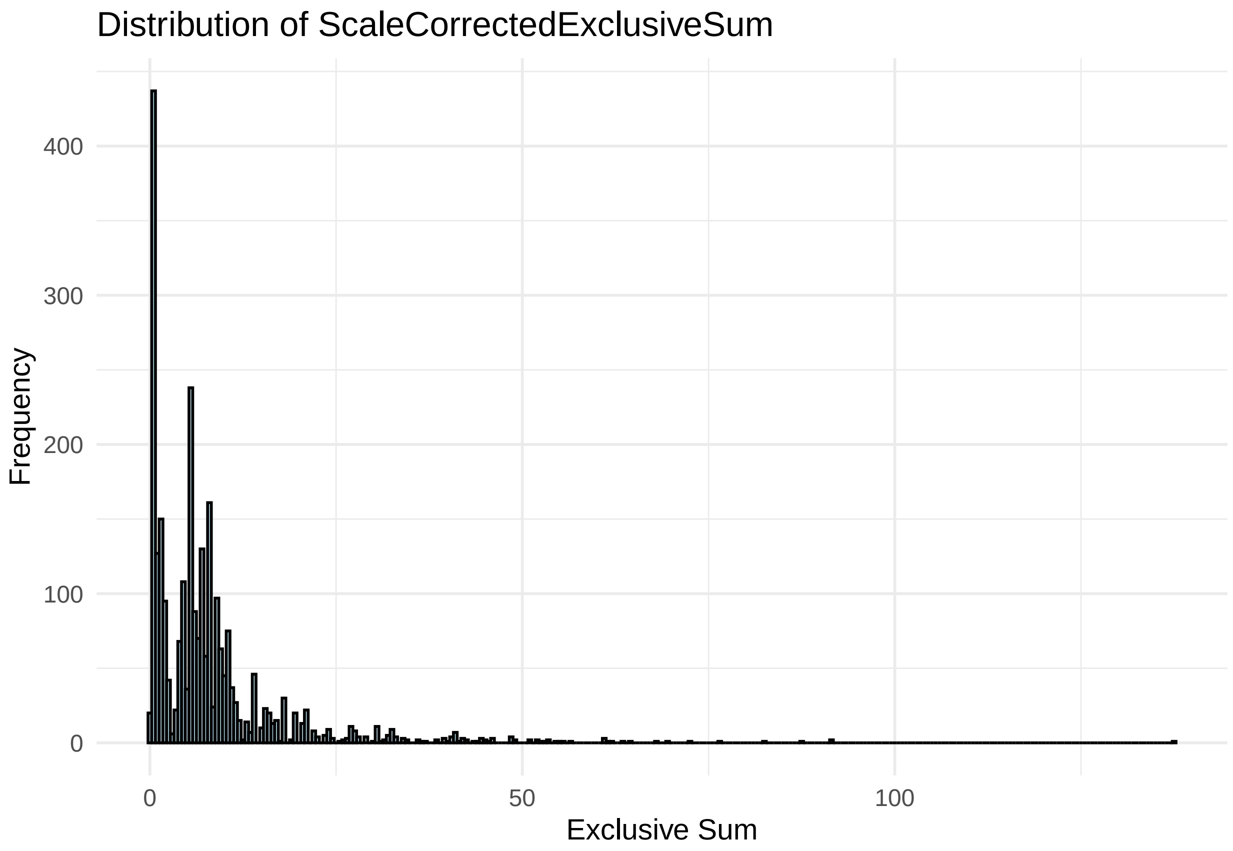
```
library(ggplot2)
library(dplyr)
VirusData <- read_csv("pt 2/VirusProteomz.csv")
```

```
## Rows: 2649 Columns: 60
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (42): MS.MS.sample.name, DepthGroup, Region, Peptide.sequence, Protein.n...
## dbl (17): z, Stn, Depth.m., Longitude, Latitude, Exclusive.Sum.PSM, Scaling_...
## lgl  (1): SOM.Label.Flag
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(VirusData)
```

```
## # A tibble: 6 x 60
##       z MS.MS.sample.name      Stn Depth.m. Longitude Latitude DepthGroup Region
##   <dbl> <chr>                <dbl>    <dbl>     <dbl>    <dbl> <chr>      <chr>
## 1     1 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## 2     2 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## 3     3 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## 4     4 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## 5     5 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## 6     6 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## # i 52 more variables: Peptide.sequence <chr>, Protein.name <chr>,
## #   Exclusive.Sum.PSM <dbl>, Scaling_Factor <dbl>,
## #   Calculated.Total.Protein..ug.L. <dbl>,
## #   Scaled.Corrected.Exclusive.Sum <dbl>, blast.accession <chr>,
## #   blast.best.hit.taxon.id <dbl>, KO <chr>, Group <chr>, Domain <chr>,
## #   Phylum <chr>, Class <chr>, Order <chr>, Family <chr>, Genus <chr>,
## #   Species <chr>, LCA.Group <chr>, LCA.Domain <chr>, LCA.Phylum <chr>, ...
```

```r
ggplot(VirusData, aes(x = Scaled.Corrected.Exclusive.Sum)) +
  geom_histogram(binwidth = 0.5, fill = "lightblue", color = "black") +
  labs(title = "Distribution of ScaleCorrectedExclusiveSum", x = "Exclusive Sum", y = "Frequency") +
  theme_minimal()
```



Boxplot: Show spread and outliers for a variable.

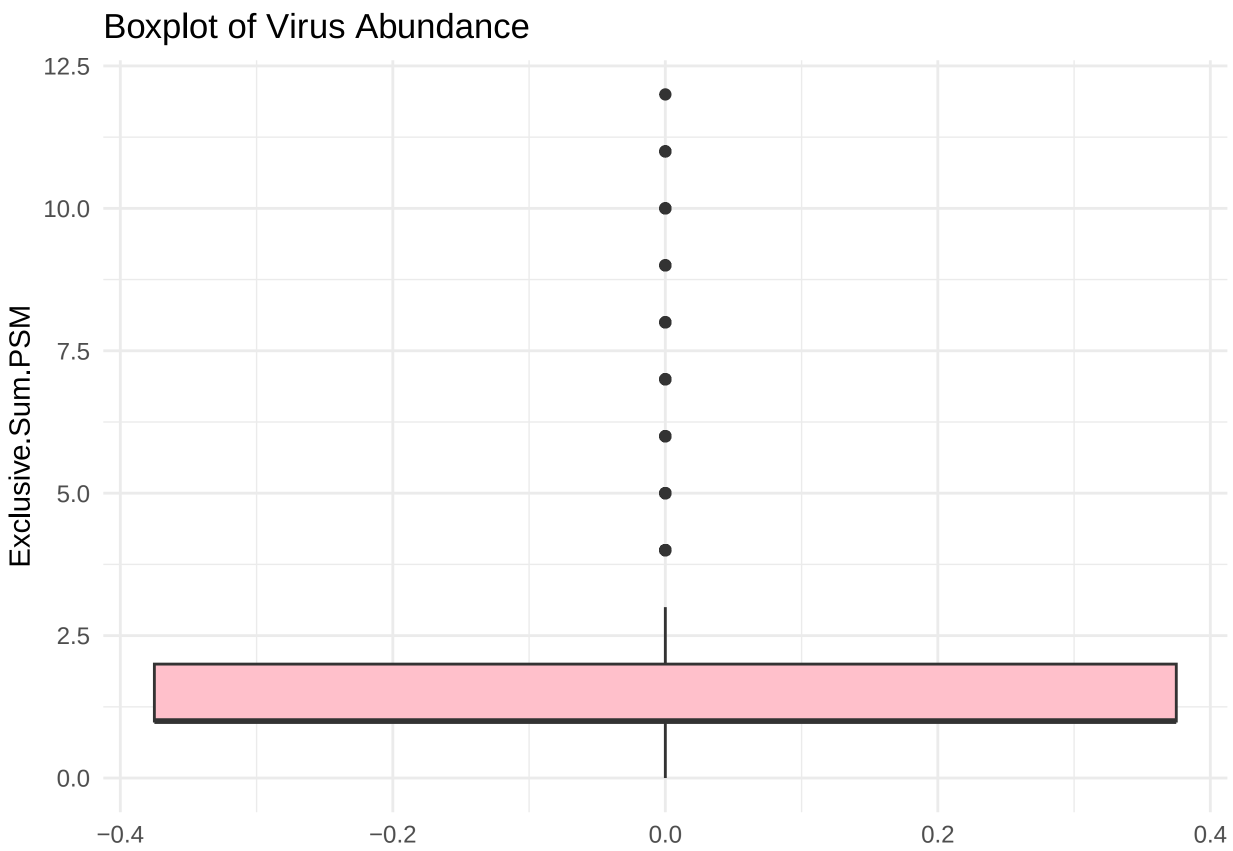
```r
library(ggplot2)
library(dplyr)
VirusData <- read_csv("pt 2/VirusProteomz.csv")
```

```
## Rows: 2649 Columns: 60
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (42): MS.MS.sample.name, DepthGroup, Region, Peptide.sequence, Protein.n...
## dbl (17): z, Stn, Depth.m., Longitude, Latitude, Exclusive.Sum.PSM, Scaling_...
## lgl  (1): SOM.Label.Flag
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
head(VirusData)
```

```
## # A tibble: 6 x 60
##       z MS.MS.sample.name      Stn Depth.m. Longitude Latitude DepthGroup Region
##   <dbl> <chr>                <dbl>    <dbl>     <dbl>    <dbl> <chr>      <chr>
## 1     1 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## 2     2 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## 3     3 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## 4     4 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## 5     5 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## 6     6 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## # i 52 more variables: Peptide.sequence <chr>, Protein.name <chr>,
## #   Exclusive.Sum.PSM <dbl>, Scaling_Factor <dbl>,
## #   Calculated.Total.Protein..ug.L. <dbl>,
## #   Scaled.Corrected.Exclusive.Sum <dbl>, blast.accession <chr>,
## #   blast.best.hit.taxon.id <dbl>, KO <chr>, Group <chr>, Domain <chr>,
## #   Phylum <chr>, Class <chr>, Order <chr>, Family <chr>, Genus <chr>,
## #   Species <chr>, LCA.Group <chr>, LCA.Domain <chr>, LCA.Phylum <chr>, ...
```

```
ggplot(VirusData, aes(y = Exclusive.Sum.PSM)) +
  geom_boxplot(fill = "pink") +
  labs(title = "Boxplot of Virus Abundance", y = "Exclusive.Sum.PSM") +
  theme_minimal()
```



# Boxplot: Show Family level abundance by Stn group.

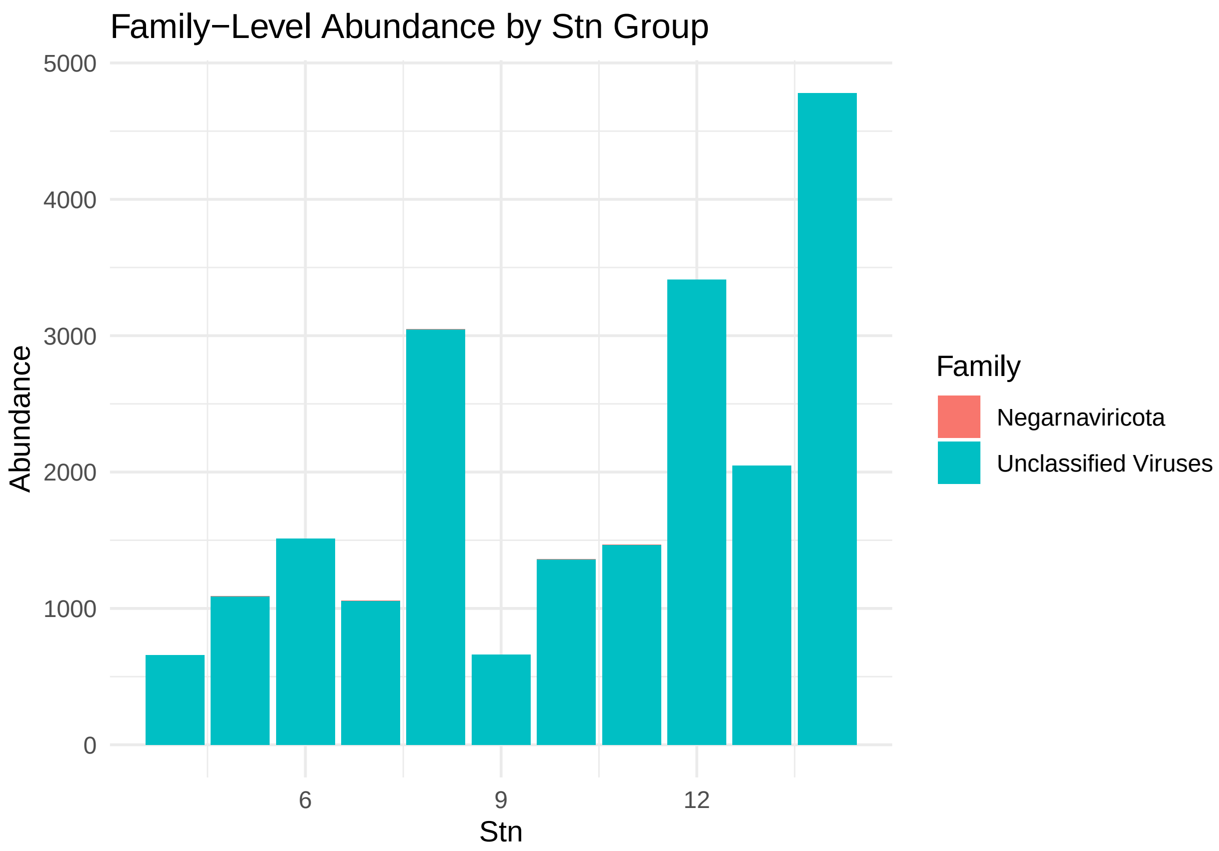
```r
library(ggplot2)
library(readr)

# Remove rows with missing values in key columns
cleaned_data <- VirusData[!is.na(VirusData$Family) & !is.na(VirusData$Stn), ]


# Aggregate data by DepthGroup and Family
taxa_abundance <- aggregate(cleaned_data$Scaled.Corrected.Exclusive.Sum,
                             by = list(Stn = cleaned_data$Stn, Phylum = cleaned_data$Phylum),
                             FUN = sum)

# Rename columns
colnames(taxa_abundance) <- c("Stn", "Family", "Abundance")

# Plot
ggplot(taxa_abundance, aes(x = Stn, y = Abundance, fill = Family)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Family-Level Abundance by Stn Group", x = "Stn", y = "Abundance") +
  theme_minimal()
```



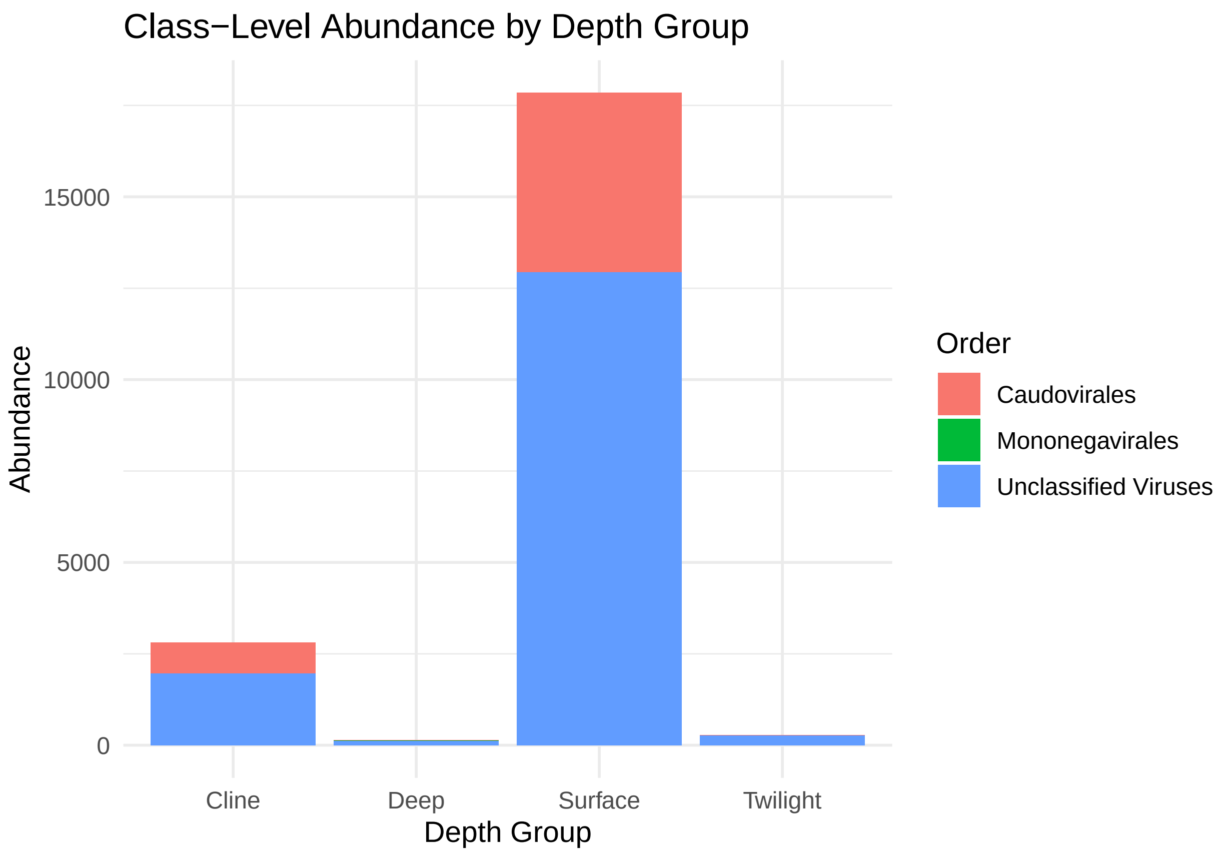
# Boxplot: Show Order level abundance by depth group.

```
library(ggplot2)
library(readr)

# Aggregate data by DepthGroup and Order
taxa_abundance_order <- aggregate(cleaned_data$Scaled.Corrected.Exclusive.Sum,
                                   by = list(DepthGroup = cleaned_data$DepthGroup, Order = cleaned_data
                                   FUN = sum)

# Rename columns
colnames(taxa_abundance_order) <- c("DepthGroup", "Order", "Abundance")

# Plot
ggplot(taxa_abundance_order, aes(x = DepthGroup, y = Abundance, fill = Order)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Class-Level Abundance by Depth Group", x = "Depth Group", y = "Abundance") +
  theme_minimal()
```



Boxplot: Show Order taxa abundance by Class group.

```
library (readr)
library (ggplot2)
VirusData <- read_csv("pt 2/VirusProteomz.csv")
```

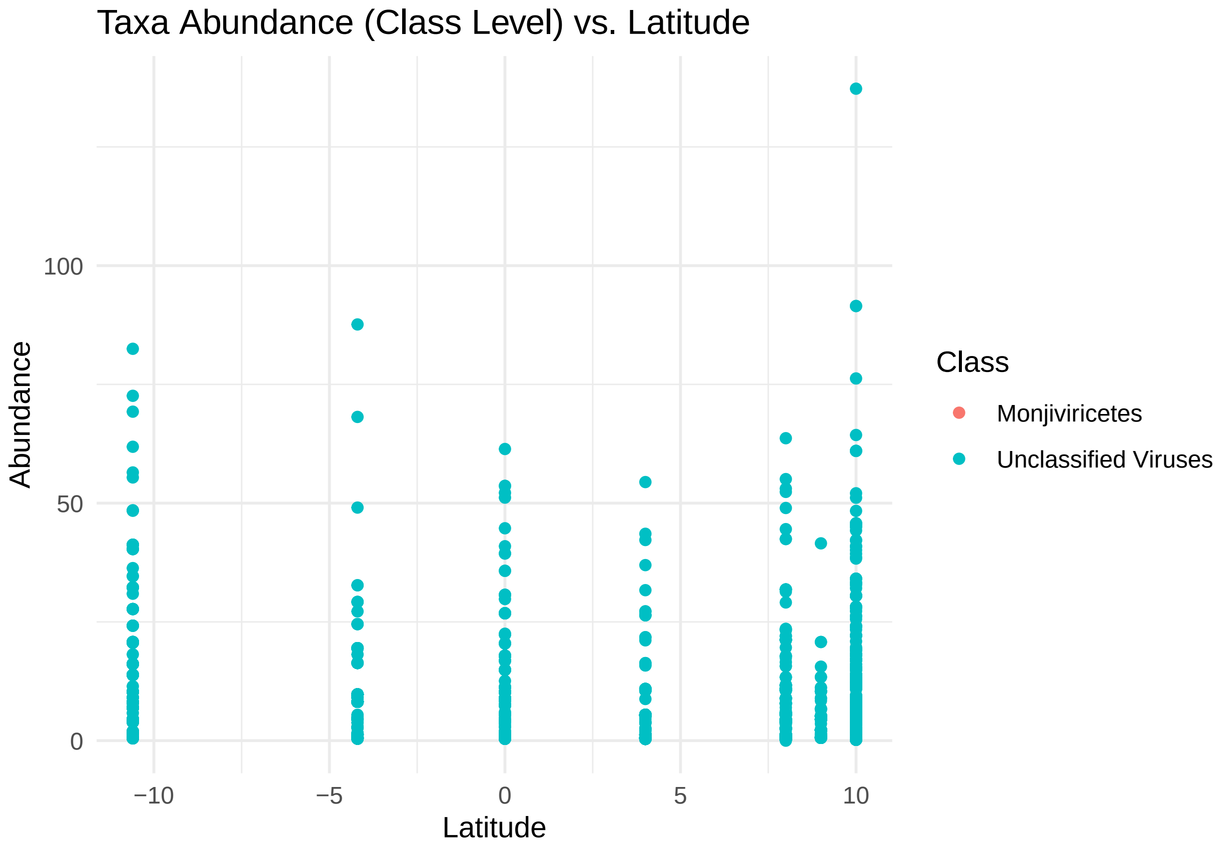
```
## Rows: 2649 Columns: 60
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (42): MS.MS.sample.name, DepthGroup, Region, Peptide.sequence, Protein.n...
## dbl (17): z, Stn, Depth.m., Longitude, Latitude, Exclusive.Sum.PSM, Scaling_...
## lgl  (1): SOM.Label.Flag
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
head(VirusData)
```

```
## # A tibble: 6 x 60
##       z MS.MS.sample.name      Stn Depth.m. Longitude Latitude DepthGroup Region
##   <dbl> <chr>                <dbl>    <dbl>     <dbl>    <dbl> <chr>      <chr>
## 1     1 170825_proteOMZ_2D_~    10       20      -140        8 Surface    South
## 2     2 170825_proteOMZ_2D_~    10       20      -140        8 Surface     South
## 3     3 170825_proteOMZ_2D_~    10       20      -140        8 Surface     South
## 4     4 170825_proteOMZ_2D_~    10       20      -140        8 Surface     South
## 5     5 170825_proteOMZ_2D_~    10       20      -140        8 Surface     South
## 6     6 170825_proteOMZ_2D_~    10       20      -140        8 Surface     South
## # i 52 more variables: Peptide.sequence <chr>, Protein.name <chr>,
## #   Exclusive.Sum.PSM <dbl>, Scaling_Factor <dbl>,
## #   Calculated.Total.Protein..ug.L. <dbl>,
## #   Scaled.Corrected.Exclusive.Sum <dbl>, blast.accession <chr>,
## #   blast.best.hit.taxon.id <dbl>, KO <chr>, Group <chr>, Domain <chr>,
## #   Phylum <chr>, Class <chr>, Order <chr>, Family <chr>, Genus <chr>,
## #   Species <chr>, LCA.Group <chr>, LCA.Domain <chr>, LCA.Phylum <chr>, ...
```

```
ggplot(cleaned_data, aes(x = Latitude, y = Scaled.Corrected.Exclusive.Sum, color = Class)) +
  geom_point() +
  labs(title = "Taxa Abundance (Class Level) vs. Latitude", x = "Latitude", y = "Abundance") +
  theme_minimal()
```

```
## Warning: Removed 24 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



Conclusion

In this notebook, we have demonstrated various R functionalities from basic operations to advanced analyses on virus data. We started with basic data structures and operations, moved on to data cleaning and exploratory data analysis, and finally performed advanced analyses such as regression and clustering. Future work could include more sophisticated modeling and validation techniques to further understand the virus data.